In [1]:
```
import numpy as np
import pandas as pd

pd.set_option('display.max_rows',None)
pd.set_option('display.max_columns', None)
```

In [2]:
```
movies=pd.read_csv("tmdb_5000_movies.csv")
credits=pd.read_csv("tmdb_5000_credits.csv")
```

In [3]:
```
movies.shape
```

Out[3]:
```
(4803, 20)
```

In [4]:
```
movies.head(2)
```

Out[4]:

	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id": 1463, "name": "culture clash"}, {"id":...	en	Avatar	In the 22nd century, a paraplegic Marine is di...	150.437577
1	300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	http://disney.go.com/disneypictures/pirates/	285	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	en	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	139.082615

In [5]:
```
credits.shape
```

Out[5]:
```
(4803, 4)
```

In [6]:
```
credits.sample(10)
```

Out[6]:

	movie_id	title	cast	crew
1658	8054	The Imaginarium of Doctor Parnassus	[{"cast_id": 4, "character": "Doctor Parnassus...	[{"credit_id": "52fe448ec3a36847f809cd2b", "de...
2259	29078	The Game of Their Lives	[{"cast_id": 1, "character": "Frank Borghi", "...	[{"credit_id": "52fe45c6c3a368484e06e337", "de...
815	8488	Hitch	[{"cast_id": 17, "character": "Alex 'Hitch' Hi...	[{"credit_id": "52fe44adc3a36847f80a3c35", "de...
3668	11935	Capricorn One	[{"cast_id": 1, "character": "Robert Caulfield...	[{"credit_id": "55fe2597c3a36813370021e6", "de...
244	254128	San Andreas	[{"cast_id": 2, "character": "Ray Gaines", "cr...	[{"credit_id": "55dd12ca9251417444000ce5", "de...
29	37724	Skyfall	[{"cast_id": 1, "character": "James Bond", "cr...	[{"credit_id": "52fe46689251416c910537ad", "de...
4634	14290	Better Luck Tomorrow	[{"cast_id": 2, "character": "Ben Manibag", "c...	[{"credit_id": "52fe45e39251416c75065da3", "de...
3607	16441	The Beastmaster	[{"cast_id": 1, "character": "Dar", "credit_id...	[{"credit_id": "52fe46d49251416c75084ec9", "de...
3518	8982	The Protector	[{"cast_id": 1, "character": "Kham", "credit_i...	[{"credit_id": "52fe44cec3a36847f80ab0c7", "de...
3374	4251	Veer-Zaara	[{"cast_id": 11, "character": "Veer Pratap Sin...	[{"credit_id": "52fe43b6c3a36847f8069a81", "de...

In [7]:
```
df= movies.merge(credits, on='title')
df.shape
```

Out[7]:
```
(4809, 23)
```

In [8]:
```
df.head(2)
```

Out[8]:

	budget	genres	homepage	id	keywords	original_language	original_title	overview	popularity
0	237000000	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	http://www.avatarmovie.com/	19995	[{"id": 1463, "name": "culture clash"}, {"id":...	en	Avatar	In the 22nd century, a paraplegic Marine is di...	150.437577
1	300000000	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	http://disney.go.com/disneypictures/pirates/	285	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	en	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	139.082615

remove columns="budget", "homepage", "original_language", "original_title", "original_title", "popularity", "production_companies", "production_countries", "release_date", "revenue", "runtime", "spoken_languages", "status", "tagline", "vote_average", "vote_count", "movie_id"

In [10]:
```
df = df[['movie_id', 'title', 'overview', 'genres', 'keywords', 'cast', 'crew']]
df.head(2)
```

Out[10]:

	movie_id	title	overview	genres	keywords	cast	crew
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...	[{"id": 28, "name": "Action"}, {"id": 12, "nam...	[{"id": 1463, "name": "culture clash"}, {"id":...	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	[{"id": 12, "name": "Adventure"}, {"id": 14, "...	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...

In [11]:
```
df.isnull().sum()
```

Out[11]:
```
movie_id    0
title       0
overview    3
genres      0
keywords    0
cast        0
crew        0
dtype: int64
```

In [12]:
```
df.dropna(inplace=True)
```

In [13]:
```
df.isnull().sum()
```

Out[13]:
```
movie_id    0
title       0
overview    0
genres      0
keywords    0
cast        0
crew        0
dtype: int64
```

In [14]:
```
df.duplicated().sum()
```

Out[14]:
```
0
```

In [15]:
```
df.iloc[0].genres
```

Out[15]:
```
'[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 878, "name": "Science Fiction"}]'
```

In [16]:
```
import ast
```

In [17]:
```
def convert(text):
    L=[]
    for i in ast.literal_eval(text):
        L.append(i['name'])
    return L
```

In [18]:
```
df['genres']=df['genres'].apply(convert)
df.head(2)
```

Out[18]:

	movie_id	title	overview	genres	keywords	cast	crew
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...	[Action, Adventure, Fantasy, Science Fiction]	[{"id": 1463, "name": "culture clash"}, {"id":...	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	[Adventure, Fantasy, Action]	[{"id": 270, "name": "ocean"}, {"id": 726, "na...	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...

In [19]:
```
df['keywords']=df['keywords'].apply(convert)
df.head(2)
```

Out[19]:

	movie_id	title	overview	genres	keywords	cast	crew
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...	[Action, Adventure, Fantasy, Science Fiction]	[culture clash, future, space war, space colon...	[{"cast_id": 242, "character": "Jake Sully", "...	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	[Adventure, Fantasy, Action]	[ocean, drug abuse, exotic island, east india ...	[{"cast_id": 4, "character": "Captain Jack Spa...	[{"credit_id": "52fe4232c3a36847f800b579", "de...

In [20]:
```
import ast
ast.literal_eval('[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventure"}, {"id": 14, "name": "Fantasy"}, {"id": 878, "n
```

Out[20]:
```
[{'id': 28, 'name': 'Action'},
 {'id': 12, 'name': 'Adventure'},
 {'id': 14, 'name': 'Fantasy'},
 {'id': 878, 'name': 'Science Fiction'}]
```

In [21]:
```
def convert3(text):
    L=[]
    counter=0
    for i in ast.literal_eval(text):
        if counter != 3:
            L.append(i['name'])
            counter+=1
        else:
            break
    return L
```

In [22]:
```
df['cast']=df['cast'].apply(convert3)
df.head(2)
```

Out[22]:

	movie_id	title	overview	genres	keywords	cast	crew
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...	[Action, Adventure, Fantasy, Science Fiction]	[culture clash, future, space war, space colon...	[Sam Worthington, Zoe Saldana, Sigourney Weaver]	[{"credit_id": "52fe48009251416c750aca23", "de...
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...	[Adventure, Fantasy, Action]	[ocean, drug abuse, exotic island, east india ...	[Johnny Depp, Orlando Bloom, Keira Knightley]	[{"credit_id": "52fe4232c3a36847f800b579", "de...

In [23]:
```
def fetch_director(text):
    L=[]
    if text is not None:  # Handle None values
        for i in ast.literal_eval(text):
            if i['job']=='Director':
                L.append(i['name'])
    return L
```

In [24]:
```
df['crew']=df['crew'].apply(fetch_director)
df.sample(2)
```

Out[24]:

	movie_id	title	overview	genres	keywords	cast	crew
1738	10317	Our Brand Is Crisis	A feature film based on the documentary "Our B...	[Comedy, Drama]	[bolivia, woman, political campaign, south ame...	[Sandra Bullock, Anthony Mackie, Billy Bob Tho...	[David Gordon Green]
1229	8457	Drillbit Taylor	Three kids hire a low-budget bodyguard to prot...	[Comedy]	[prison, jealousy, homeless person, beach, par...	[Owen Wilson, Leslie Mann, Josh Peck]	[Steven Brill]

In [25]:
```
df['overview']=df['overview'].apply(lambda x:x.split())
df.head(2)
```

Out[25]:

	movie_id	title	overview	genres	keywords	cast	crew
0	19995	Avatar	[In, the, 22nd, century,, a, paraplegic, Marin...	[Action, Adventure, Fantasy, Science Fiction]	[culture clash, future, space war, space colon...	[Sam Worthington, Zoe Saldana, Sigourney Weaver]	[James Cameron]
1	285	Pirates of the Caribbean: At World's End	[Captain, Barbossa,, long, believed, to, be, d...	[Adventure, Fantasy, Action]	[ocean, drug abuse, exotic island, east india ...	[Johnny Depp, Orlando Bloom, Keira Knightley]	[Gore Verbinski]

In [26]:
```
df['genres']=df['genres'].apply(lambda x:[i.replace(" ","") for i in x])
df['keywords']=df['keywords'].apply(lambda x:[i.replace(" ","") for i in x])
df['cast']=df['cast'].apply(lambda x:[i.replace(" ","") for i in x])
df['crew']=df['crew'].apply(lambda x:[i.replace(" ","") for i in x])
df.head()
```

Out[26]:

	movie_id	title	overview	genres	keywords	cast	crew
0	19995	Avatar	[In, the, 22nd, century,, a, paraplegic, Marin...	[Action, Adventure, Fantasy, ScienceFiction]	[cultureclash, future, spacewar, spacecolony, ...	[SamWorthington, ZoeSaldana, SigourneyWeaver]	[JamesCameron]
1	285	Pirates of the Caribbean: At World's End	[Captain, Barbossa,, long, believed, to, be, d...	[Adventure, Fantasy, Action]	[ocean, drugabuse, exoticisland, eastindiatrad...	[JohnnyDepp, OrlandoBloom, KeiraKnightley]	[GoreVerbinski]
2	206647	Spectre	[A, cryptic, message, from, Bond’s, past, send...	[Action, Adventure, Crime]	[spy, basedonnovel, secretagent, sequel, mi6, ...	[DanielCraig, ChristophWaltz, LéaSeydoux]	[SamMendes]
3	49026	The Dark Knight Rises	[Following, the, death, of, District, Attorney...	[Action, Crime, Drama, Thriller]	[dccomics, crimefighter, terrorist, secretiden...	[ChristianBale, MichaelCaine, GaryOldman]	[ChristopherNolan]
4	49529	John Carter	[John, Carter, is, a, war-weary,, former, mili...	[Action, Adventure, ScienceFiction]	[basedonnovel, mars, medallion, spacetravel, p...	[TaylorKitsch, LynnCollins, SamanthaMorton]	[AndrewStanton]

In [27]:
```
df['tag']=df['overview'] + df['genres'] + df['keywords'] + df['cast'] + df['crew']
```

In [28]:
```
updated_df= df[['movie_id','title','tag']]
updated_df.head(2)
```

Out[28]:

	movie_id	title	tag
0	19995	Avatar	[In, the, 22nd, century,, a, paraplegic, Marin...
1	285	Pirates of the Caribbean: At World's End	[Captain, Barbossa,, long, believed, to, be, d...

In [29]:
```
updated_df['tag'] = updated_df['tag'].apply(lambda x:" ".join(x))
updated_df.head(2)
```

```
/var/folders/gl/xrbhkv5j2rb0pw50rbz6n8lc0000gn/T/ipykernel_91695/1049375847.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  updated_df['tag'] = updated_df['tag'].apply(lambda x:" ".join(x))
```

Out[29]:

	movie_id	title	tag
0	19995	Avatar	In the 22nd century, a paraplegic Marine is di...
1	285	Pirates of the Caribbean: At World's End	Captain Barbossa, long believed to be dead, ha...

In [30]:
```
updated_df['tag'] = updated_df['tag'].apply(lambda x:x.lower())
updated_df['tag'][1]
```

```
/var/folders/gl/xrbhkv5j2rb0pw50rbz6n8lc0000gn/T/ipykernel_91695/2775826312.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  updated_df['tag'] = updated_df['tag'].apply(lambda x:x.lower())
```

Out[30]:
```
"captain barbossa, long believed to be dead, has come back to life and is headed to the edge of the earth with will turner and elizabeth swann. but nothing is quite as it seems. adventure fantasy action ocean drugabuse exoticisland eastindiatradingcompany loveofone'slife traitor shipwreck strongwoman ship alliance calypso afterlife fighter pirate swashbuckler aftercreditsstinger johnnydepp orlandobloom keiraknightley goreverbinski"
```

Stop words removal

In [32]:
```
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer(max_features=5000, stop_words='english')
```

Stemming

Stemming is a text preprocessing technique to reduce words to their root or base form. The goal of stemming is to simplify and standardize words, which helps improve the performance of information retrieval, text classification, and other NLP tasks.

In [35]:
```
! pip install nltk
```

```
Requirement already satisfied: nltk in /opt/anaconda3/lib/python3.12/site-packages (3.9.1)
Requirement already satisfied: click in /opt/anaconda3/lib/python3.12/site-packages (from nltk) (8.1.7)
Requirement already satisfied: joblib in /opt/anaconda3/lib/python3.12/site-packages (from nltk) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in /opt/anaconda3/lib/python3.12/site-packages (from nltk) (2024.9.11)
Requirement already satisfied: tqdm in /opt/anaconda3/lib/python3.12/site-packages (from nltk) (4.66.5)
```

In [36]:
```
from nltk.stem.porter import PorterStemmer
ps = PorterStemmer()
```

In [37]:
```
def stem(text):
    y=[]
    for i in text.split():
        y.append(ps.stem(i))
    return " ".join(y)
```

In [38]:
```
updated_df['tag']=updated_df['tag'].apply(stem)
```

```
/var/folders/gl/xrbhkv5j2rb0pw50rbz6n8lc0000gn/T/ipykernel_91695/3063456617.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  updated_df['tag']=updated_df['tag'].apply(stem)
```

In [39]:
```
vectors = cv.fit_transform(updated_df['tag']).toarray()
vectors
```

Out[39]:
```
array([[0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       ...,
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0],
       [0, 0, 0, ..., 0, 0, 0]])
```

In [40]:
```
vectors[0]
```

Out[40]:
```
array([0, 0, 0, ..., 0, 0, 0])
```

In [41]:
```
cv.get_feature_names_out()
```

Out[41]:
```
array(['000', '007', '10', ..., 'zone', 'zoo', 'zooeydeschanel'],
      dtype=object)
```

cosine similarity

In [43]:
```
from sklearn.metrics.pairwise import cosine_similarity

similarity=cosine_similarity(vectors)
similarity[0]
```

Out[43]:
```
array([1.        , 0.08346223, 0.0860309 , ..., 0.04499213, 0.        ,
       0.        ])
```

In [44]:
```
sorted(list(enumerate(similarity[0])), reverse=True, key=lambda x:x[1])[1:6]
```

Out[44]:
```
[(1214, 0.28676966733820225),
 (2405, 0.26901379342448517),
 (3728, 0.2605130246476754),
 (507, 0.255608593705383),
 (539, 0.25038669783359574)]
```

In [45]:
```
def recommend(movie):
    movie_index=updated_df[updated_df['title']==movie].index[0]
    distances=similarity[movie_index]
    movies_list=sorted(list(enumerate(distances)), reverse=True, key=lambda x:x[1])[1:6]
    for i in movies_list:
        print(df.iloc[i[0]].title)
```

In [46]:
```
recommend('Batman Begins')
```

```
The Dark Knight
Batman
Batman
The Dark Knight Rises
10th & Wolf
```

In [47]:
```
import pickle
```

In [48]:
```
# pickle.dump(updated_df,open('movie_list.pkl','wb'))
# pickle.dump(similarity,open('similarity.pkl','wb'))
```

In [49]:
```
pickle.dump(updated_df.to_dict(),open('movie_dict.pkl','wb'))
```

In [50]:
```
pickle.dump(similarity, open('similarity.pkl','wb'))
```

In [ ]: